Chapter One

Introduction

On September 23, 1999, NASA fired rockets that were intended to put its Mars Climate spacecraft into a stable low-altitude orbit over the planet. But after the rockets were fired, the spacecraft disappeared—scientists speculated that it had either crashed on the Martian surface or had escaped the planet completely. This disaster was the result of confusion over measurement units—the manufacturer of the spacecraft had specified the rocket thrust in pounds, whereas NASA assumed that the thrust had been specified in metric system newtons (Browne 2001).

Although measurement is important in the physical sciences, it is equally important in the social sciences, including the discipline of criminology. Criminologists, policy makers, and the general public are concerned about the levels of crime in society, and the media frequently report on the extent and nature of crime. These media reports typically rely on official data and victimization studies and often focus on whether crime is increasing or decreasing.

Both official crime and victimization data indicated that property and violent crime in the United States were in a state of relatively steady decline from the early 1990s to 2000. But in late May 2001, the release of Federal Bureau of Investigation (FBI) official crime data, widely publicized in the media, indicated that crime was no longer declining. This prompted newspaper headlines such as "Decade-Long Crime Drop Ends" (Lichtblau 2001a) and led commentators such as James Alan Fox, dean of the College of Criminal Justice at Northeastern University, to assert, "It seems that the crime drop is officially over. . . . We have finally squeezed all the air out of

the balloon" (as quoted in Butterfield 2001a). However, some two weeks after the release of these official data, a report based on victimization data indicated that violent crime had decreased by 15 percent between 1999 and 2000, the largest one-year decrease since the federal government began collecting victimization data in 1973 (Rennison 2001). This prompted headlines such as "Crime Is: Up? Down? Who Knows? (Lichtblau 2001b) and led James Alan Fox to declare, "This is good news, but it's not great news" (as quoted in Bendavid 2001).

How do we reconcile the conflicting messages regarding crime trends from these two sources? First, although most media sources commenting on the FBI data failed to mention this caveat, the official data report was in fact based on preliminary data: "[The report] does not contain official figures for crime rates in 2000" (Butterfield 2001a). Second, and more important, the underlying reason for these differences is that the two data sources measured crime differently. Official crime data are based on reports submitted to the FBI by police departments and measure homicide, rape, robbery, aggravated assault, burglary, car theft, and larceny. Victimization data are based on household surveys that question respondents about their experiences with crime and do not include homicide (for obvious reasons). However, the victimization survey does include questions about simple assaults, which are far more common than aggravated assaults or robberies and thus tend to statistically dominate the report. As Butterfield (2001b) pointed out, simple assaults accounted for 61.5 percent of all violent crimes identified in the victimization survey, and because they had declined by 14.4 percent in 2000 compared with 1999, they accounted for most of the decline in violent crime revealed in the victimization data. In short, and as Alfred Blumstein (as quoted in Butterfield, 2001b) noted, "[The data] are telling us that crime is very difficult to measure."

Statistics and numerical counts of social phenomena, including crime, have become a major fact of modern life. Countries are compared and ranked in terms of statistical information on health, education, social welfare, and economic development. Cities and individuals are compared on similar kinds of social indicators. Geographical areas, social groups, and individuals are judged as relatively high, low, or normal on the basis of various numerical counts. Notice the variability in the following numerical data from a statistical profile of the United States (U.S. Bureau of the Census 1999).

- There were 41,518 injuries associated with a hammer in 1997. There were 44,335 injuries from toilets and 37,401 injuries from televisions in the same year.

- Whooping cough deaths increased from 1,700 to 7,400 from 1980 to 1998. Deaths in the United States resulting from gonorrhea decreased from 100,400 to 35,600 in the same time frame.
- More than 54 percent of U.S. citizens in 1997 were classified as overweight, and 19 percent were considered obese.
- Broccoli consumption increased from 1.4 pounds per capita in 1980 to 5.6 pounds per capita in 1998.
- The proportion of households with a television set increased from 2 percent in 1965 to 85 percent in 1998.
- There were 14 deaths per 100,000 population in the United States by firearms in 1995 compared to 4 deaths from falling per 100,000 population. In contrast, the death rate by firearms in England and Wales was only 0.4 per 100,000, but the death rate from falling was 26 per 100,000.
- The number of offenses known to the police per 100,000 population was 8,836 in the District of Columbia in 1998 and 4,071 in Montana in the same year.

At their face value, each of these types of statistical information may serve as a basis for social action. For example, this information may lead people to exhibit more care when using hammers or toilets, have greater concern with their nagging cough and less concern about particular sexually transmitted diseases, feel good or bad about their weight, invest in broccoli and television production, watch out for firearms in the United States and hazardous walking conditions in Great Britain, and relocate to Montana. It is also not uncommon for this type of numerical data to form the basis of public policy. In fact, public health programs, law enforcement, and other agencies rely on these descriptive statistics to implement various types of reform.

Before taking corrective actions based on statistical information, however, it is important to consider several questions about its accuracy and how the data were collected. These questions about the measurement of social phenomena are often neglected in public discourse, but they ultimately will determine whether corrective action is necessary. For example, one's opinion about the statistics presented above may change when one considers the following questions regarding the measurement of these social facts:

- How are injuries by hammers, toilets, and televisions counted? If a television repairperson hits a television with a hammer and it falls in the toilet and results in an electric shock to the repairperson, is this classified as an injury by a hammer, toilet, or television? Do all agencies classify these injuries the same way? Note that these injury data are calculated from a sample of hospitals with emergency treatment departments. If people are injured by these products and do not go to an emergency ward, their injuries will not be counted. Under these conditions, the number of injuries by hammers, toilets, or televisions may be substantially higher or even lower, depending on how they are counted.

- Does the rise in whooping cough deaths reflect an actual increase in these fatalities or is it due to medical advances in the last decade that have now made it easier to detect whooping cough as a medical problem? Is the dramatic decline in gonorrhea deaths due to improvements in medical care and early detection of this disease or is it due to the reclassification of sexually transmitted disease (STD) deaths by medical personnel (e.g., some STD deaths are now attributed to AIDS)?

- Who establishes the categories of overweight and obesity? Are these self- reported feelings (e.g., "Do you consider yourself to be overweight or obese?"), or are they standardized by height and bone structure? Are the figures of 54 percent overweight and 19 percent obese derived from the entire U.S. population or particular subsets of individuals who went for some type of medical treatment (e.g., diabetes or heart disease, which are associated with obesity)?

- Are we really measuring changes in the human consumption of broccoli or the amount of broccoli purchased per capita? For example, the increase in the last two decades in the number of exotic pets (like iguanas) that eat broccoli may artificially inflate the estimates of human consumption. How are the figures for broccoli consumers who grow their own broccoli counted in data that are derived from grocery stores? Can an increase in a small number of "super" broccoli eaters underlie this increase instead of the apparent rise in the proportion of consumers over time?

- Are uniform standards for the measurement of firearm and falling deaths used within and across the United States and Great Britain? If a person dies a year after the injury occurred, how is the death classified? How are multiple causes counted (e.g., a person is shot and then falls off a building, or a person falls onto a loaded gun)?

- Counts of crime in the District of Columbia include offenses reported to the police at the National Zoo, thereby inflating the crime rate per 100,000 residents. In contrast, complete police data in Montana were not available, so the crime rates had to be estimated. If different jurisdictions use different rules for counting crime and citizens report crime differently in rural and urban areas (which they do), conclusions about the relative danger of the District of Columbia and the safety of Montana may be premature.

As these examples illustrate, numerical measures of crime and other social phenomena have enormous potential to inform social scientists about their theories of human behavior, provide politicians and legislators with an empirical basis for public policy decisions, and help the general public structure their routine activities and how they live their lives. Unfortunately, many people who use these statistics are grossly uninformed about how they are collected, what they mean, and their strengths and limitations.

The goal of this book is to critically examine the various ways in which crime is measured and thereby to instill a healthy skepticism about the accuracy of current methods of counting crime. All social measurement involves human decisions, interpretations, and errors. By examining the sources of error in the measurement of crime, social scientists, legislators, and the general public will be in a better position to understand the utility of current theory and crime control practices that derive from statistical data on crime. In later chapters, we address in considerable detail issues surrounding the three most commonly used measures of crime and delinquency: official data, self-report, and victimization studies. In this introductory chapter, we address the measurement of social phenomena in the context of the key concepts of reliability and validity.

Reliability, Validity, and Sources of Error in the Measurement of Social Phenomena

Stevens (1959) defined measurement as "the assignment of numerals to events or objects according to rule" (p. 25). The initial steps in measure-

ment are to (1) clarify the concept one is interested in and (2) construct what is known as an operational definition of that concept. An individual's social class is often operationally defined by income level, educational attainment is usually measured by years of formal schooling, sexual promiscuity is gauged by number of sexual partners, and political party preference is measured by one's expressed attitudes toward Democrats and Republicans. As illustrated by these examples, the process of operationalization and measurement involves the attachment of a specific meaning to abstract concepts.

The accuracy of many measures of social phenomena, however, is both context and time specific. Sexual promiscuity, for example, was judged by different standards in the Victorian period of the 1800s, the "free love" era of the 1960s, and the current period. Similarly, our working definitions of crime are context and historically specific. Prostitution, alcohol use, and drug use may be differentially evaluated as "serious" crime, depending on the geographic location and historical period, the political circumstances, and the prevailing legal structures. Although illegal in most of the United States, prostitution is legal in certain jurisdictions in the state of Nevada. The consumption and sale of alcohol are legal in the present-day United States, but they were illegal in the 1920s. And although some of the most severe penalties in our criminal code are reserved for users of substances such as cocaine, marijuana, heroin, and methamphetamine, these substances were not illegal in the United States prior to the 20th century. Under these conditions, our choice of a particular working definition and unambiguous indicator of a concept becomes more difficult.

Selecting precise indicators of abstract concepts is a crucial step in attempting to operationalize any social phenomena. Within this process, two fundamental properties of good measurement exist: reliability and validity. Reliability is concerned with questions related to the stability and consistency of measurement over repeat trials, and validity refers to the extent of congruence between the operational definition and the concept it purports to measure.

Reliability and validity are easily demonstrated when we consider the measurement of intelligence. If a test of intelligence sometimes yields a high intelligence quotient (IQ) and at other times a low IQ for the same individual, the test would be considered unreliable because it failed to achieve consistent results over repeated trials. An intelligence test would have questionable validity if there were differences in its ability to accurately measure the intellectual capacity of individuals from different cultures or races or both. In fact, one of the major criticisms of standard intelligence tests is

their low validity because they are not culturally sensitive (i.e., the test does not measure intelligence but instead indicates one's adaptation to middle-class culture). Although a valid measure can be unreliable, a reliable measure is not necessarily valid (e.g., a thermometer is a reliable measure of temperature but an invalid measure of social class).

Reliability and Validity in Survey Research

Many of the social measures and indicators we discuss in this chapter and two of the most frequently used measures of crime and delinquency—self-report and victimization studies—rely on surveys of various segments of the general public to collect data and to construct measures. A number of issues related to survey methodology encourage caution in interpreting the results of studies employing this methodology. These include problems in sampling and response rates to surveys, questionnaire format and wording, and interviewer effects.

Survey methodology is based on probability sampling theory. The basic principle is that a randomly selected, relatively small percentage of a population can be used to represent the attitudes, opinions, or behaviors of all people in the population if the sample is selected correctly. The key to being able to generalize to the larger population from a smaller sample is related to a fundamental principle in sampling theory known as equal probability of selection. This simply means that each member of the population has an equal, or at least known, chance of being chosen to participate in the survey. It is instructive to discuss the principles of probability sampling in the context of the frequent public opinion polls conducted in the United States by organizations such as Gallup and Roper.

In telephone surveys conducted by such organizations, the usual goal is to generalize the results of the survey to all adults, 18 years of age and older, living within the continental United States (Newport, Saad, and Moore 1997). However, such surveys generally do not cover individuals living in institutions, including college students who live on campus, armed forces personnel living on military bases, prisoners, hospital patients, and others living in group settings or housing. The procedure organizations such as Gallup use is to obtain a computerized list of all telephone exchanges in the United States, accompanied by estimates of the number of residential households attached to those exchanges. Then, through a procedure known as random digit dialing (RDD), a computer is used to generate a list of telephone numbers. This random digit dialing procedure is important in the context of obtaining a representative sample, because without it, the esti-

mated 30 percent of households in the United States that have unlisted phone numbers would not be included in the sampling frame.

The typical sample size for public opinion polls is between 1,000 and 1,500 respondents. However, the actual number of people interviewed in a survey is much less important than adherence to the equal probability of selection principle. As Newport et al. (1997) note, if respondents are not selected according to equal probability of selection principles, it would be possible to conduct a survey with a million people that could turn out to be less representative of the population than a survey conducted with only 1,000 individuals.

The accuracy of estimates derived from these samples is also based on probability theory. With the typical sample size of 1,000, the results are highly likely to accurately represent the true population value within a margin of error of plus or minus three percentage points. For example, the results of a Gallup poll released in May of 1998 indicated that 64 percent of the U.S. public were familiar with the erectile dysfunction drug Viagra, which had been placed on the market only a few months earlier. This survey also revealed that 13 percent of the men interviewed indicated they would like to try the drug within the next year. Interestingly, 15 percent of the women answered that they would like their husband to try Viagra in the next year (Saad 1998). The margin of error indicates that the true rating of women who would like their husbands to try Viagra was somewhere between 12 and 18 percent. If the sample size for this survey was increased to 2,000, the results would be accurate within plus or minus two percentage points of the true population value, but the cost of conducting the survey would double.

Another important issue in assessing the reliability and validity of survey results is related to rates of response—what is also referred to as "contact and cooperation" (Singer and Presser 1989): the correspondence between the sample elements selected and those actually interviewed. In recent years, survey researchers have become concerned about the phenomenon of declining response rates to surveys, which can result in biased samples and thereby inaccurate measures or estimates. At least part of the reason for the general public's lack of willingness to participate in survey research is the proliferation of entities, both private and government, engaged in survey research. For example, the number of telemarketing firms increased from 30,000 in 1985 to more than 600,000 in 1995, and according to industry sources, more than 25 million solicitation calls are made in a single day (Bearden 1998). According to a 1994 study, one in three potential respondents refuses to participate in a survey, and even for respondents who do

participate in surveys occasionally, 38 percent had refused to participate in at least one in the previous year. More generally, it is estimated that from 1990 to 2000, the response rate to telephone surveys declined from approximately 40 percent to 15 percent (Lewis 2000). In most cases, data resulting from surveys with poor response rates can be assumed to be unrepresentative and biased, because the respondents are likely to be self-selected and different in a number of unknown ways from those who do not respond. Unfortunately, many researchers take whatever data they collect, analyze it, and derive conclusions without any consideration of the issue of nonresponse bias. A prime example of the problems that can result from inattention to issues of nonresponse bias occurred in 1985, when the Committee on Health and Long-Term Care issued a report that referred to the abuse of elderly persons in the United States as "a national disgrace." This report cited research claiming that an estimated 4 percent, or 1 million elderly persons, were victims of abuse each year. However, this estimate was based on a survey of 433 elderly residents of Washington, D.C., of whom only 73, or 16 percent of the original sample, responded. Three of these 73 respondents, representing 4.1 percent, reported experiencing some form of psychological, physical, or material abuse. The report then extrapolated from this small and undoubtedly unrepresentative sample to assert that 1 million elderly people were victims of abuse, "thereby constructing a national epidemic out of these three incidents" (Gilbert 1997:112).

The Census Bureau has a high level of respect and is admired for the quality of its data collection policies and procedures. Census Bureau staff are well trained, many of the leading experts in research methodology have direct contact with the national agency, sampling designs are among the most sophisticated in the world, statisticians that work with Census staff possess state-of-the-art knowledge about population estimation, and rigorous pretesting is conducted before actual data collection begins. But even the census, conducted every 10 years in the United States, which is intended to represent a full enumeration of the population, is subject to nonresponse bias and other problems in counting the population.[1] In 1970, the first year that government officials administered the initial part of the census by mail, 83 percent of households returned the questionnaire. In 1980, the rate of return had declined to 75 percent, and by 1990, it was only 65 percent. For the 2000 census, 67 percent of the households that received the form returned it (Holmes 2000). More important, these rates of response vary across geographical regions of the United States and across different sociodemographic categories of the population. Due to nonresponse bias and other problems in enumerating the entire population, it is estimated that

the 2000 census did not count between 1.6 and 2.7 percent of black residents and between 2.2 and 3.5 percent of Hispanics. A further 2.8 to 6.7 percent of Native Americans living on reservations were also not counted (Holmes 2001). Interestingly, the population of one town in rural Pennsylvania was missed entirely in the 2000 census. The 14 people who live in the town of Slovenska Nardona Podporna Jednota apparently were not around when the census taker visited—they thought she would come back, but she did not. As a result, the town's population for the year 2000 is listed as zero (*New York Times* 2001). In total, it is estimated that the 2000 census did not count between 6.4 and 8.6 million people living in the United States.

The reverse problem with census data is that of overcounting. It was estimated that more than 4 million people were in fact counted twice in the 2000 census (Holmes 2001). Those who are counted twice tend to be children of divorced parents, college students living away from home who independently fill out census forms but are also listed by their parents, and people with two homes who receive forms in the mail at both of their dwellings. This potentially large overcount is also related to the fact that for the 2000 census, forms were available at convenience stores and government agencies, and respondents were able to provide information over the telephone. The Census Bureau also engaged in a $102 million dollar prime-time advertising campaign to prompt individuals to participate. The issues associated with an accurate enumeration of the population are by no means trivial, because census data are used to determine how seats in the U.S. House of Representatives will be apportioned, to draw congressional and state legislative district boundaries, to allocate state and federal funds, to formulate a wide array of public policies, and to assist with planning and decision making in the private sector.

Reliability and Validity Issues Related to the Questionnaire and Respondents

A number of factors related to the survey instrument itself and the individuals responding to survey questions affect the reliability and validity of results from this method of data collection. Three of these will be covered here: question wording effects, question order effects, and response effects.

Question Wording Effects

A study of wording effects using data from the General Social Survey compared two different versions of questions on government spending priorities and revealed systematic differences in responses. When respondents were asked if they supported increased spending on "welfare," only 32 percent answered in the affirmative. However, when respondents were asked whether there should be "more assistance for the poor," 62 percent favored increased spending (Smith 1989). Another example of the effects of question wording and response options comes from studies examining support for capital punishment in the United States. An opinion poll conducted by Gallup in February of 2001 found that 67 percent of the U.S. population favored capital punishment. However, when interviewers asked whether the penalty for murder should be execution *or* life in prison with no possibility of parole, support for capital punishment declined to 54 percent (Jones 2001).

Question Order Effects

The order in which questions are asked also can have an impact on responses. For example, in a poll conducted before the 2000 U.S. presidential election to determine the popularity of candidates Gore and Bush, respondents were asked to state their preference for president after having responded to a question that asked them to evaluate then President Clinton "as a person." This ordering of questions resulted in a lower level of support for Gore, probably because the question about Clinton reminded respondents of the Monica Lewinsky scandal and led them to disapprove of his vice president as well. However, when the company conducting the poll reordered the questions and surveyed a new sample, support for Gore increased (Harwood and Crossen 2000).

Response Effects

Data from the 2000 census are also relevant to the issue of response effects. One of the most important characteristics of the U.S. population that the census attempts to measure accurately is its racial composition.[2] Although race is a social construct, the racial composition of various jurisdictions in the United States has important implications for economic and social policies. The 2000 census was the first in which people in the United States were allowed to identify themselves as belonging to more than one racial group; the six racial categories created a total of 63 possible racial combinations for respondents to self-identify. Results from the 2000 census indicate that fully 6.8 million people identified themselves as multiracial, and although 93 percent of these classified themselves into only two racial categories, 823 respondents actually checked all six racial categories

(Kasindorf and El Nasser 2001). With respect to the same question, people who indicated that they were "some other race" were asked to write in a particular race. Answers included Bolivian, Bushwacker, Cosmopolitan, and Aryan (Scott 2001).

The American Indian category offers an interesting glimpse into the complications created by the change in census racial classifications. The number of American Indians and Alaska natives who defined themselves only by that category increased by 26 percent between 1990 and 2000. However, when the number of people who claimed they were part Indian is added, the total increased to 4.1 million, representing a 110 percent increase in the number of American Indians since 1990 (Schmitt 2001). However, it is not clear that all of those who identified themselves as Native American legitimately fall into that category. An informal survey conducted by a newspaper in Spokane, Washington, for example, found that some individuals marked the Native American category "as a way to tell the U.S. Census Bureau to mind its own business." Others apparently identified themselves as Native American "because they were born in the United States" (McDonald 2001). More important, racial composition data from the 2000 census will not be directly comparable with previous census figures, and the ability to track the progress of racial groups with respect to their educational, occupational, health, and income characteristics will become far more problematic.

Although it may seem straightforward, even the classification of gender in a census can be ambiguous. In Canada, a transsexual person refused to answer the question, "Are you male or female," on that country's 2001 census. This individual, who was born a male but was taking hormones and had breasts and male genitals, noted that "my gender was not listed" (Raphael 2001).

A related problem has characterized the U.S. census with respect to identifying the number of households occupied by gay couples. In 1990, a person who shared a household with an individual of the same sex and also reported being married created a problem for census data-coders because the Census Bureau did not recognize same-sex marriages. To make the responses consistent, the Census Bureau changed either the person's sex or his or her relationship to the other person, because "if they said they were married and had a spouse of the same sex, the simple thing was to change the spouse's sex. We made them a married couple" (Spencer, quoted in Peterson, 2001). At least partially as a result of changes in this procedure in the 2000 census such that gay and lesbian householders could claim an unmarried partner and then identify his or her sex, there was a "huge increase" (Peterson 2001) in the numbers of gay households identified in 2000.

Errors in questionnaire data are also associated with response styles—the tendency to choose a certain category when responding to a question—regardless of the content of the item. For example, in the frequently used "agree-disagree" format on questionnaires, some respondents may be characterized by an acquiescence response set: the tendency to agree with a question, regardless of its content (Singleton and Straits 1999). A second response style is referred to as social desirability: the tendency to choose those response options most favorable to an individual's self-esteem or in accord with prevailing social norms, regardless of one's real position on the given question. Some have argued that social desirability effects may explain why comparisons of survey data over time reveal a general decline in overt expressions of racially prejudiced attitudes (Quillian 1996).

Additional response problems are related to issues of memory, and in this context, two types of errors can be distinguished: forgetting and telescoping in time. With respect to telescoping, events and behaviors are reported as having happened more recently than they actually did. This form of response error is particularly relevant in the context of self-report and victimization surveys, which are addressed in Chapters 3 and 4 of this book.

The very real possibility also exists that respondents, for a number of different reasons, may be somewhat less than truthful in responding to questionnaires; the evidence regarding lying on questionnaires is well documented. In a 1950 study, Parry and Crossley asked individuals a number of questions in situations where the accuracy of their answers could be assessed. The proportion of honest answers ranged from 98 percent on a question asking whether the respondents had a telephone to approximately 50 percent on one that asked about their voting behavior. McCord (1951) similarly demonstrated that people will sometimes lie when they are asked questions about things that do not exist: one third of his sample claimed they had voted in a special election that in fact never was held. Studies also suggest that between 33 and 45 percent of respondents will lie when they are asked about their level of education, and about half when they are asked whether they have received welfare assistance (Nettler 1978). In addition, some studies have suggested that the tendency to be less than truthful may vary according to the racial/ethnic and gender characteristics of respondents (Mensch and Kandel 1988).

Some surveys of criminal behavior and drug use, which will be addressed in more detail in Chapter 4, have discovered that minority groups have a greater tendency to underreport these behaviors. One explanation of this tendency is that minorities feel more threatened or are made uneasy when

asked to report on delinquent activities. Whatever the possible reasons for this underreporting, researchers conducting studies and those reporting on the results of such studies need to be aware of the possibility of biases resulting from these tendencies.

A more general concern with respect to survey research is related to respondents' general knowledge. Public opinion polls have shown that many people in the United States are unaware that there are three branches of government in the United States; significant numbers of the U.S. population believe that Brazil is the capital of Ohio, and approximately 18 percent believe that the sun circles the earth (*USA Today* 1997). In the 1989 General Social Survey, 61 percent of respondents did not feel they were able to rank the social standing of "Wisians." However, 39 percent were able to rank this group, and they provided Wisians with a rather low average rating of 4.12 on a 9-point social ranking scale (*Seattle Post-Intelligencer* 1992). Wisians were a fictitious ethnic group, added by the designers of the General Social Survey to test the honesty of respondents in answering questions.

In short, all data derived from survey research are subject to reliability and validity problems. An intelligent consumer of such data will pay attention to these issues before uncritically accepting the findings from survey research.

Measuring Crime and Deviance

We now move on to a consideration of issues that are more directly relevant to the main topic of this book: the measurement of crime and deviant behavior. We begin with a discussion of the problems associated with measuring crime on college campuses, followed by a consideration of how questionable measures of the extent of drug consumption have been used to create alleged drug "epidemics" with resulting policy changes.

Measuring College Campus Crime

Since the 1990s, numerous states and the federal government have enacted laws requiring colleges and universities to publish crime statistics. The first federal law related to this requirement, known as the Crime Awareness and Campus Security Act, was passed in 1990 (Port and Lesser 1999). As is often the case with legislative proposals in the United States, this law was enacted primarily in response to the occurrence of a single

event: the murder of 19-year-old Jeanne Clery at Lehigh University in Pennsylvania in 1986. Clery was a freshman who was assaulted and murdered while asleep in her residence room. When Clery's parents investigated the situation, they discovered that Lehigh University had not informed students about 38 violent crimes that had been committed on the campus in the three years prior to their daughter's murder. The Clerys joined with other campus crime victims and persuaded Congress to enact legislation requiring all colleges and universities to publish statistics on the amount and type of crime on campuses.

As a result of subsequent amendments to this legislation in 1998, institutions must report the incidence of homicide, manslaughter, arson, rape, robbery, aggravated assault, burglary, motor vehicle theft, drug offenses, liquor law violations, and illegal weapons possession. In addition, institutions are required to provide greater detail regarding alleged hate crimes, defined by federal law as incidents that "manifest evidence of prejudice based on race, religion, sexual orientation, or ethnicity." Campuses that do not comply with the legislation face the possibility of fines of up to $25,000 and of losing federal student aid. When data on college crime were first released in the early 1990s, several media outlets invoked rather alarmist language to describe the situation. For example, *U.S. News and World Report* (1994), commenting on the 1993 statistics, alleged that there was an "epidemic" of college campus crime. Similarly, *USA Today* (Henry 1996) referred to "steep increases in crime" in describing the 1994 campus crime statistics. But serious crime on college campuses is exceedingly rare when compared to overall crime rates in the United States—there is less than one homicide for every million students on campus in any given year in the United States.

Problems in the reliability and validity of campus crime data became apparent soon after the federal legislation was enacted. These problems ranged from confusion surrounding how to code particular crimes to outright manipulation of the statistics. A study conducted by the National Center for Education Statistics found that 40 percent of the colleges and universities were using federal definitions of crime to classify their data, 45 percent were using state definitions, and 15 percent were using definitions of their own design (Port and Lesser 1999). A 1997 audit conducted by the U.S. General Accounting Office discovered that only 2 of the 25 colleges examined were correctly reporting their crime statistics. Among other omissions, some colleges were routinely omitting rapes and other sexual assaults that were reported to school officials but not to the police. For example, in September of 1999, the University of Florida admitted with-

holding 35 rapes from its annual crime reports for the years 1996, 1997, and 1998. Instead of the 12 rapes that were recorded in the official report for this period, the university was aware of 47; however, university officials claimed that they believed that rapes reported to a victims' advocacy group should not be counted (Port and Lesser 1999).

Perhaps the most notorious example of the manipulation of campus crime statistics occurred at the University of Pennsylvania. In 1996, this university reported 18 robberies in its federally mandated campus security report, whereas the police blotter indicated that 181 robberies had occurred. The apparent reason for this gross discrepancy was that the university had chosen to exclude crimes that had occurred on sidewalks and streets that crossed the campus and in buildings it did not own (Port and Lesser 1999).

Anomalies in the officially recorded data and incidents such as the one that occurred at the University of Pennsylvania resulted in further amendments to the legislation. Beginning in 1998, institutions were required to report crimes occurring on public property that was "reasonably contiguous" to their campuses. Not surprisingly, there was initially considerable confusion on the part of university officials regarding what constituted reasonably contiguous property; it has since been defined as public sidewalks, streets, and parking lots adjacent to a campus, or any public property running through the campus.

Comparisons of crime data across college campuses in the United States suggest that universities are not adopting the same definitions of contiguous areas, however. For example, campus police at the University of Washington in Seattle expressed skepticism when the 1998 figures on campus crime were released. In that year, the University of Southern California, located in the middle of a high-crime area of South Central Los Angeles, recorded only 4 assaults, whereas the University of Washington recorded 93 (Rivera 2000). In 1999, the University of Washington's 127 drug arrests placed it fourth in the nation. However, campus police noted that the arrest sometimes involved street people and individuals who wandered onto the campus (Rivera 2001). The perils associated with uninformed comparisons of these data are also revealed when we consider the situation of colleges and universities with branch campuses. The 1997 report for the University of Idaho, located in a rural area of the state, indicated that seven rapes had occurred on campus in that year. However, the rapes had actually occurred at a smaller branch campus of the university, located in Coeur d'Alene. Similarly, Eastern Washington University, located in a largely rural area of Washington state, recorded 74 aggravated assaults in 1997, but the overwhelming majority of these had occurred in a contiguous area of the university's branch campus in the heart of downtown Spokane (deLeon and Sudermann 2000).

Two additional categories of campus crime to examine are those of alcohol and drug arrests. Between 1997 and 1998, alcohol arrests on college campuses increased by 24.3 percent nationally, whereas arrests for violations of drug legislation increased by 11.1 percent. However, campus law enforcement officials attribute these increases to tougher enforcement of existing drug and alcohol guidelines and changes in the previously mentioned reporting categories stipulating that colleges had to include crimes taking place in reasonably contiguous areas.

At the University of Wisconsin, where arrests for alcohol increased from 342 in 1997 to 792 in 1998, the campus police chief claimed that the 132 percent change was due to the university's hiring more campus police officers who were more vigorous in enforcing the laws. At the University of North Carolina at Greensboro, which experienced more than a 700 percent increase in drug arrests between 1997 and 1998, the increases were attributed to the expanded geographical area for which crimes were recorded; of the 132 drug arrests in 1998, 88 occurred on public property near the campus and 17 in residence halls, areas the college had not included in its 1997 report (Nicklin 2000).

There has also been considerable confusion regarding the procedures for counting these drug and alcohol arrests. The University of New Hampshire at Durham was unable to meet the Department of Education's reporting deadline of October 24 for their 1997 and 1998 data. When officials at the university asked the Department of Education how to deal with this problem, they were told to record no offenses for these categories. As a result, an uninformed perusal of the "official data" for the University of New Hampshire would lead one to believe that this campus had no drug arrests in 1997 and 1998 and 124 in 1999, instead of what actually occurred—56 arrests in 1997 and 85 in 1998 (Nicklin 2000).

In addition to the problems outlined above with respect to counting drug and alcohol crimes or offenses, stipulations in the legislation requiring institutions to report the number of campus disciplinary referrals for violations of alcohol, drug, and weapons law violations have created further confusion. In the 1998 report, several institutions placed arrests and referrals in the same category, creating the illusion of a significant increase in arrests. For example, Wake Forest University reported an increase from 8 to 298 for alcohol-related arrests between 1997 and 1998; however, officials at the university claimed they had made only one liquor arrest—the remaining 297 were referrals (Nicklin 2000).

Given all the problems associated with the collection and coding of these data, it makes little sense to engage in cross-campus and over-time comparisons of the campus crime data.

The CAP Index

An alternative measure and ranking of college campuses with respect to their levels of crime has been created by a risk assessment company. This CAP (crimes against persons) index focuses on crime risk in neighborhoods surrounding college campuses and estimates "the risk of crime for the coming year through a sophisticated computer model that compares socioeconomic data to past reports of actual crime" (Port and Lesser 1999). Publication of these CAP index rankings of college campuses has attracted considerable controversy, especially in light of the fact that in 1999, four historically black colleges appeared in the top five most dangerous list and seven historically black colleges were rated in the top ten (Wright 2000). In fact, four of these institutions were located in the same urban neighborhood in Atlanta, and officials from the colleges concerned expressed confusion over the rankings. For example, although it experienced no murders, no sexual assaults, only 6 simple assaults and 17 robberies on its campus from 1996 to 1998 inclusive—hardly indicative of a high level of dangerousness—Morehouse College was ranked as the fifth most dangerous campus in the United States, according to this index. In response to criticism of these rankings, the creators of the CAP index claim that the method is 70 to 90 percent accurate in predicting actual levels of crime. However, the company refused to reveal the precise statistical methods used in creating the index—"that would be like Coca-Cola giving away its formula for Coke" was the claim of the chief executive officer of the company. Without the ability to independently verify the reliability and validity of the various components of this index, however, the rankings that result from it must be treated with skepticism.

Drugs and "Drug Epidemics"

Illegal drugs have been a major concern of policy makers in the United States since the beginning of the 20th century. And, as is the case in other areas of social, economic, and crime policies, competing interests rely on both official and unofficial data to support their respective agendas.

Prior to the 1996 presidential election, incumbent President Bill Clinton presented data from victimization surveys to suggest that there had been a 9

percent decrease in violent crime in the United States and claimed that the decline was due to the effectiveness of his administration's crime policies. Republican candidate Bob Dole saw things differently, and used self-report data from the Federal Department of Health and Human Services to blame Clinton for a doubling of drug use among teenagers. However, the questions used in the 1994 survey that led Dole to attack Clinton were very different from those used in previous drug surveys, and the agency could not ensure that it had successfully adjusted for the differences. Even more important, many of the increases in drug use to which Dole referred were not statistically significant. Heroin use by teenagers, for example, superficially doubled from 0.3 percent in 1994 to 0.7 percent in 1995, but the actual number of users in the sample of 4,600 surveyed had increased from only 14 to 32 (Schoor 1996).

An additional example of the confusion that can be caused by uninformed comparisons of drug use statistics comes from the 1999 Report of the Office of National Drug Control Policy. That report claimed that there were 1.5 million people in the United States who had used cocaine in the previous month. However, the same document claimed that 3.6 million people in the United States had used cocaine in the past week (Caulkins 2000). Clearly, these estimates are highly inconsistent and difficult to reconcile. The explanation for the large discrepancy in these estimates is that the first was based exclusively on data from the National Household Survey on Drug Abuse, whereas the latter included data from the Drug Use Forecasting program, which collects self-reports of drug use among arrestees in local jails, who are more likely to use drugs.

Questionable official and unofficial data on drug use are frequently used to justify changes in drug policies. An interesting example of this phenomenon occurred in 2000 and 2001, when the popular media published hundreds of articles on an alleged epidemic in the use of the drug ecstasy (MDMA). A March 5, 2001 editorial, written by former federal drug czar William Bennett (2001), claimed that "while the crack cocaine epidemic of the 1990s has passed, methamphetamine and ecstasy are growing in popularity, especially among the young." Bennett did not provide statistics, official or otherwise, to support his claim of this increase in the use of ecstasy. However, a survey that was widely cited in the media, conducted under the auspices of the Partnership for a Drug Free America, reported that the percentage of teenagers using ecstasy had doubled between 1995 and 2000—from 5 to 10 percent.

Given the paucity of additional self-report data on the use of ecstasy, especially by adults, media sources relied on alternative measures, such as

reported seizures of ecstasy tablets, reports of law enforcement officials, and emergency room admission data, to support their claim of an "alarming explosion" (Rashbaum 2000) in the use of MDMA. The commissioner of the U.S. Customs Service claimed that seizures of ecstasy by his agency had increased from 350,000 pills in 1997 to 3.5 million in 1999, then to 2.9 million in just the first two months of 2000. He projected that seizures would amount to 7 or 8 million by the end of 2000. An *Associated Press* article (Hays 2000) suggested that "seizures of the tablets . . . have multiplied like rabbits." An article in *USA Today* (2001a) noted that "ecstasy, a drug once used primarily at nightclubs, has expanded beyond the club scene and is being sold at high schools, on the street, and even at coffee shops in some cities." The source of these claims of ecstasy use spreading to previously unknown contexts was an informal convenience survey of officials in 20 cities in the United States, 80 percent of whom said that ecstasy was "more available than ever."

An additional measure of the alleged increase in ecstasy use came from the federal Drug Abuse Warning Network (DAWN), which tracks hospital room emergency admissions. Rashbaum (2000) reported that mentions of the drug in this source increased from 68 in 1993 to 637 in 1997 (the latest year for which statistics were available).

Despite the questionable validity of the statistics used to document this ecstasy epidemic, in March of 2001, the U.S. Sentencing Commission enacted harsh new penalties for MDMA. These penalties treat ecstasy offenders more severely than cocaine offenders, resulting in a 5-year sentence for individuals selling 200 grams (approximately 800 pills) of the substance and a 10-year sentence for those selling 2,000 grams or more (Lindesmith Center 2001). These legislative changes were enacted despite the opposition of many medical experts and researchers, who argued that the substance was far less likely to cause violence than drugs such as alcohol and was less addictive than cocaine or tobacco. Advocates of the increased penalties argued that these were necessary to curb ecstasy use by teenagers and young adults (*Washington Post* 2001).

Apparently, ecstasy has become a growing problem in Canada as well. In May of 2000, a drug enforcement officer from Toronto claimed, "I believe ecstasy has reached epidemic proportions in this country" (as quoted in Godfrey 2000). Given similar problems with respect to the availability of current statistics on the actual extent of ecstasy use, the Canadian media also relied extensively on seizure figures to support the claim that ecstasy use had increased. In an article appearing in the *National Post*, Grey (2000) reported that seizures of ecstasy in Canada had doubled between 1998 and

1999. Police across the country seized 712,000 ecstasy tablets in 1999, with an estimated street value of between $17.8 and $28.5 million. The article also claimed that it was becoming "common knowledge" among law enforcement officials and researchers that ecstasy was "the drug of choice across demographic lines." In May of 2000, several Canadian newspapers announced that the largest seizure of ecstasy in Canadian history had taken place at Pearson International Airport in Toronto. Police reported that they had seized 170,000 ecstasy tablets, valued at $5 million. However, it turned out that police had made a mathematical error in their calculations, weighing the quantity of pills per pound instead of per kilogram. Thus, the actual seizure was 61,000 tablets, valued at $1.8 million. Ben Soave, a superintendent for the Royal Canadian Mounted Police, noted, "It's one of those unfortunate situations. It was an error that we made and we're only human. So I apologize for that" (as quoted in Alphonso 2000). The ecstasy problem was given further publicity when testimony given at an inquest into the death of a Toronto youth alleged that 13 deaths had been caused by the substance during a three-year period beginning in 1998. Although these ecstasy-related deaths were widely published in the media, it was eventually determined that seven of the deaths were the result of individuals using drug "cocktails," mixtures of heroin, cocaine, and methadone (Freed 2000). Although as of March 2001, no federal or provincial legislation had been enacted in Canada to deal with the ecstasy "problem," a "Raves Act" for the city of Toronto was proposed in May of 2000. This legislation would have defined a rave as a dance event occurring between 2:00 a.m. and 6:00 a.m. for which admission was charged. It would have increased police powers of arrest in situations where drugs were sold at such events and allowed them to terminate the event if illegal acts were occurring (Freed 2000). We need to question whether it is good public policy to change policies based on such questionable data.

To conclude, it is clear that the data we have addressed in this chapter are subject to reliability and validity problems and are also subject to varying interpretations. And although such data are frequently used to draw attention to social problems and issues, to theoretically explain the causes of these problems, and to influence policies to deal with them, it is important to remain critical of the construction of these measures. As Campbell (1971, as cited in Johnston and Carley 1981) argued, "The more any social indicator is used for decision-making, the more subject it will be to corrupting pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." More specific to crime data, Gurr (1977) notes that in the interpretation of crime data, it is necessary to "dis-

entangle the social reality of behavioral change from the political and administrative reality of change in the institutions which respond to and record behavioral change" (p. 117).

The Design of This Book

How we measure, but too often mismeasure, crime and delinquency/deviance is the general topic of this book. This chapter introduced the primary constraints on constructing useful measurements of any social phenomena while illustrating those constraints with specific examples from research efforts on crime and deviance. Chapter 2 offers a historical overview of the measurement of crime, paying particular attention to the evolution from official data to self-report studies to victimization surveys. We see that many of the shortcomings characteristic of contemporary data on crime were identified by social scientists writing in the late 1800s and early to mid-1900s. Chapter 3 focuses on official measures of crime based on data compiled by local, state, and federal law enforcement agencies and examines problems associated with the collection and interpretation of these data. Chapter 4 reviews self-reported measures of criminal as well as deviant activity and, in the process, provides an overview of the methodology—survey research—used to collect both self-report and victimization data on crime. Chapter 5 describes victimization surveys in detail, highlighting the advantages as well as the disadvantages of measuring crime on the basis of victim accounts. The final chapter in this book moves us from the realm of describing various measures (and mismeasures) of crime and delinquency/deviance to the province of applying these measures to particular situations. Here we see that accurate and appropriate measurement is absolutely essential in testing explanations or theories of criminal behavior and in developing crime prevention or control policies.

Notes

1. Although the task of enumerating the entire population of a country such as the United States is monumental, consider the situation of China. In that country, with an estimated population of 1.3 billion, the census is conducted through face-to-face interviews by more than 10 million volunteers and government workers. In the past, social and political issues have affected the accuracy of population counts in China. Demographers estimate that between 20 million and 100 million people are left out of the count due to evasion of household taxes and child-bearing restrictions, which limit urban families to a single child and rural families to two children if their firstborn is a girl (Chang 2000).
2. In the first U.S. census, conducted in 1790, blacks were counted as three fifths of a person and Native Indians were not counted at all (Anderson and Fienberg 1999).